DOCUMENT RESUME

ED 160 610 TM 007 316

AUTHOR Anderson, Beverly L.
TITLE Evaluation of the Michigan Educational Assessment Program Using the Shepard Evaluation Model.
PUB DATE Mar 78
NOTE 29p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Toronto, Ontario, Canada, March, 1978)

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS *Check Lists; *Educational Assessment; Educational Programs; Elementary Secondary Education; *Evaluation Criteria; *Evaluation Methods; Formative Evaluation; Models; *Program Evaluation; *State Programs; Testing Programs
IDENTIFIERS *Michigan Educational Assessment Program; Shepard (Loretta); *Shepard Evaluation Model

ABSTRACT

In spring 1977, a five-member team used the Shepard evaluation method to review the Michigan Educational Assessment Program (MEAP). The team's activities were the initial stage of a complete formative MEAP assessment. The Shepard evaluation method consists of a checklist which provides nearly 50 topics to investigate when conducting an evaluation of an assessment program. A sample modified checklist is included. The topics in the checklist are grouped into five areas: goals and purposes; technical aspects; management; intended and unintended effects; and costs. The Shepard method also entails the following evaluation activities: purpose and planning; staffing the evaluation; data collection; data sources; management, and data synthesis and reporting. The team attempted to determine the extent to which program staff had sufficient information to make decisions in the areas included in the checklist. A list of questions derived from the checklist for use in the MEAP study is appended. The team members found the Shepard method to be very useful in structuring an evaluation. However, they felt that management, data synthesis, and reporting areas required more elaboration and emphasis than was found in the Shepard Model. (Author/JAC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED160610

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

EVALUATION OF THE MICHIGAN EDUCATIONAL ASSESSMENT PROGRAM USING THE SHEPARD EVALUATION MODEL

Beverly L. Anderson
Northwest Regional Educational Laboratory
Portland, Oregon

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Northwest Regional Educational Lab.

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

Paper presented at the National Council of Measurement in Education Convention, Toronto, Ontario, March 1978.

TM007 316

INTRODUCTION

In March 1977 representatives of the Michigan Educational Assessment Program (MEAP) and the National Assessment of Educational Progress (NAEP) met with Loretta Shepard, author of a method of evaluating assessments,[1] to make plans for an evaluation of MEAP.[2] NAEP was interested in having the Shepard evaluation method used to evaluate a statewide assessment program and MEAP was interested in having such an evaluation conducted. Thus the two organizations chose to jointly sponsor the work of an external panel which would apply the Shepard evaluation method to MEAP.

At the March meeting a five-member panel was identified according to guidelines presented in the Shepard evaluation method. Each panel member was to devote five days to the task (with the team leader having an additional five days). The Shepard checklist, which provides nearly 50 topics to investigate when doing an evaluation of an assessment program, was to guide the work of the panel. The topics in the Shepard checklist are grouped into five areas: (a) goals and purposes, (b) technical aspects, (c) management, (d) intended and unintended effects, and (3) costs.

On April 27, 1977, Judith Moyer of MEAP, Jack Schmidt of NAEP and Loretta Shepard met with the panel in Chicago to clarify the panel's responsibilities and operating procedures. Given the desire to have the panel's work completed by June, it was necessary to limit the scope or procedures of the study. The panel, jointly with MEAP, agreed that all five areas of the checklist needed to be considered. And, given the time

---

[1] Shepard, L. A. A method of evaluating assessment. Paper presented at the Sixth Annual Conference on Large Scale Assessment, Boulder, Colorado, June 1976.

[2] Appendix A contains a summary of the MEAP prepared by an MEAP staff member.

constraints, it was agreed that the most useful information could be obtained if the panel focused on *the extent to which MEAP has and is gathering useful information to serve as the basis for decisions in each of the above five areas*. It was further decided that the panel's approach would be to review existing documents and interview key decision makers and MEAP staff to see if adequate information exists. The panel was not commissioned to generate new data, e.g., about the technical quality of the tests or the program effects, but rather to judge whether or not MEAP has adequate information available to make decisions regarding the topics presented in the checklist. Thus, the panel did not actually conduct an evaluation of MEAP.

At the April meeting the panel also reviewed the Shepard checklist and developed specific questions to guide their review. It was necessary to continually keep in mind that the primary task had been defined as that of determining if sufficient evidence existed regarding the topics. Each question was therefore phrased in terms of "Does sufficient evidence exist to judge the adequacy of. . .".

Using the list of questions developed by the panel, the MEAP staff sent each panelist relevant documents to review during May. After reviewing documents, each panelist identified people (by name or position) whom they wished to interview. The MEAP contact person made arrangements for these interviews. The panel members did not attempt to interview people beyond MEAP staff, legislators, news media representatives, Michigan Education Association (MEA) personnel and key Michigan Department of Education (MDE) decision makers, since they were focusing only on the extent to which evidence now is available to MEAP staff to have an evaluation of the program done.

From June 1 to 3 panelists conducted interviews in Lansing with MEAP staff, legislators, news media personnel, MEA staff and key MDE decision makers and wrote an initial draft of their report. On June 16 the team leader make a presentation at the Seventh Annual Conference on Large Scale Assessment on the panel's response to the use of the Shepard method and some findings of the panel.[3] In July, the panel's final report was submitted to the MDE. Subsequently, MEAP prepared a written response to the report. Both the panel's report and the MEAP response are available during this session.[4]

In this presentation I will critique the Shepard evaluation method. I will leave to the MDE representative the task of discussing the specifics of the panel's observations and how Michigan is responding to them.

REVIEW OF SHEPARD METHOD[5]

In her paper, Shepard identifies her purpose in presenting a method of evaluation as the following: "To prompt evaluators to ask questions and seek effects in areas they might otherwise have missed" (p. 1). She also states a secondary purpose: "Knowing what an evaluation of an assessment program ought to entail should be helpful to assessment staff. Not only could they more effectively marshal their arguments and documentation, but the evaluation plan in this paper could foster self-study" (p. 2).

---

[3] Anderson, B. L. Is Michigan assessment as a. good, b. bad--as everyone says? A tryout of the Shepard Evaluation Model. Paper presented at the Seventh Annual Conference on Large Scale Assessment, Boulder, Colorado, June 1977.

[4] Copies can also be obtained by writing to the Michigan Department of Education.

[5] The remainder of this paper is taken largely from the author's paper referenced in footnote 3.

The checklist of topics to consider when doing an evaluation of an assessment program is the focal point of the paper. Shepard also addresses these activities:

Staffing the Evaluation -- The primary issue discussed is the need for diversity among the team of evaluators and the need for objectivity or at least prior identification of bias among team members. She states: "A method of evaluation has little validity if the conclusions it generates depend on the evaluators selected" (p. 10).

Purpose of the Evaluation -- "Specifying the purposes of an evaluation makes it more likely that the purposes will be accomplished. . . The purposes of the evaluation will determine how the evaluation is conducted, what information is collected and how it is reported. . . . One of the most important clarifications of purpose is the formative-summative distinction" (p. 12).

Strategies for Data Collection -- In addition to discussing the use of hearings, transcript review and interviews, Shepard states: "Evaluations of assessment will necessarily draw heavily on opinion surveys. This is risky. Opinions may be fairly accurate reflections of feelings about an assessment but may not be very good indicators of other kinds of effects. . . .Evaluators will have to plan strategies so as not to miss effects. . . .Evaluators should collect information by more than one method to balance the errors of a single method. . . .In addition, the evaluation design should be flexible enough to allow for the identification of additional data sources to corroborate the data collected in the first round" (pp. 14-16).

Description of the Assessment -- "Evaluation reports should include a narration about the following aspects of assessment:

- History
- Rationale and Purpose
- Development
- Procedures
- Dissemination of Results" (p. 16)

Modifying the Checklist -- "The best way to use the Shepard Checklist is to review the categories and specify what each means in a specific assessment context. Some subcategories may be omitted if they do not apply. However, these same categories may be reinstated if early data collection suggests they are relevant after all. The categories of my checklist are not discrete. They interact and give each other meaning" (p. 19).

Synthesizing Checklist Findings

At the conclusion of the paper, Shepard briefly discusses synthesis of data. She makes the following statement: "The checklist produces a collection of judgments--B+ on sampling plan, C- on public knowledge about results. But, these are raw judgments. In the end, the evaluator must help combine the information so that the assessment can be judged. Unfortunately, a simple gradepoint average will not do. Some of the checklist items are more important than others. Some must be combined or weighed one against the other" (p. 44).

Shepard presents many more useful suggestions for the evaluation of an assessment. I would encourage you to read or reread her paper.

CRITIQUE OF THE SHEPARD METHOD[6]

My general response to the Shepard model is very positive. The Checklist was helpful in our initial planning session and little of what Shepard said is contrary to our experience in Michigan. My major contention is that much more could be said to give guidance to anyone--an external evaluation team, an assessment program staff, a legislative committee, a superintendent of public instruction--who plans to initiate an evaluation of an assessment program. Thus, my comments expand on the Shepard method, particularly the activities other than the use of the Checklist. I will discuss these activities in addition to looking at the Checklist:

- Purpose and planning
- Staffing the evaluation
- Data collection
- Data sources
- Management
- Data synthesis and reporting

Since the heart of Shepard's paper deals with the Checklist, I will begin by looking at it.

Modified Checklist

As Shepard mentioned in her paper, the evaluators will need to review the checklist categories and decide what each means in a specific assessment context. Such an activity was the major work of the panel at our April 27, 1977 meeting. Since we were going to be looking at the extent to which evidence already exists on the quality of checklist categories, we began by going through the checklist and developing specific questions related to the categories. None of the categories were

[6]Throughout my discussion I am going to be assuming that formative evaluation is being done at the request of those responsible for the assessment. I will leave it to someone else to discuss cases of summative evaluation or situations where an outside opponent of assessment is evaluating a program. Some differences will surely exist.

eliminated at that time but a few were added. Table 1 presents the Shepard Checklist with our additions in italics. I believe our additions are self-explanatory. Notice that we did not make changes at the broadest level of the checklist, i.e., the five areas of (a) goals and purposes, (b) technical aspects, (c) management, (d) intended and unintended effects, and (e) costs. Other ways of organizing the checklist do, of course, exist but we found this one satisfactory.

Appendix B contains our translation of the checklist categories into specific questions to direct our work.[7] It was necessary to continually keep in mind that our primary task had been defined as that of determining if sufficient evidence existed regarding the topics. Thus, each question is phrased in terms of "Does sufficient evidence exist to judge the adequacy of . . .".

This step of modifying the checklist and developing specific questions is a critical one since it determines the direction the evaluation effort will take. It is essential that the phrasing of the specific questions be consistent with the purpose of the evaluation.

If an evaluation team were actually judging the quality of each category in the checklist, the questions in Appendix B could be phrased in terms of "Does evidence indicate adequacy of . . ." Such a rephrasing markedly changes the data collection and evaluation activities in which the team would engage. Thus, the step of modifying the checklist means much more than adding or eliminating categories.

---

[7] Most of these questions are ones we developed at our planning meeting. A few were added as we reviewed documents and interviewed individuals. Others were added at the end of the evaluation as we thought of questions which still need to be addressed by MEAP.

TABLE 1

Modified Shepard's Checklist for Evaluating Assessment*

A. Goals and Purposes

Kinds of Goals:

- Pupil diagnosis
- Pupil certification
- Program evaluation
  - program improvement
  - program selection
- Resource allocation
- Accountability
- Research
- Informing the public

Criteria for Judging Goals:

- Importance
- Uniqueness
- Feasibility
- Scope

Process for Goals Development

B. Technical Aspects

Tests:

- Criterion-referenced vs. norm-referenced
- Content validity
- Cultural bias
- Field testing *and revision*
- *Development procedures*

Sampling

Administration

Data analysis:

- *For test revision*
- *For reporting*

Reporting:

- (Data analysis)**
- Different reports for different audiences
- Interpretations
- *Credibility*
- *Pervasiveness (dissemination)*

C. Management

Planning

Documentation of process

Formative evaluation

Personnel

Equal opportunity

Support Services

---

*Items in italics were added based on the Michigan experience.

**Data analysis was made a more major category.

Subcontractors

Effective decision-making procedures *and communication mechanisms*

Redress of grievances

Timelines

Division of Responsibilities

Time allocations

Relationship to other programs

Advisory groups

D. Intended and Unintended Effects

People and Groups Who May Be Affected:

- Students
- Teachers
- Parents
- Principals
- District superintendents
- District curriculum specialists
- Regional personnel
- State curriculum experts
- State legislators
- Taxpayers
- Employers
- Special interest groups
- Researchers
- *Administration organizations*
- *Key communicators at local and other levels*
- *Curriculum organizations*
- *Other state departments of education*
- *News media*
- *Teacher organizations*
- *PTA's*
- *Federal legislators and legislators in other states*

Kinds of Effects:

- Outcome and process effects
- Immediate and long-term effects
- *Direct and indirect effects*
- *Knowledge*
- Decisions
- Laws
- Resource allocation
- Effects of assessment technology
- Opinions

Criteria for Judging Effects:

- *Need*
- *Market*
- *Consistency with goals*
- *Appropriateness of use*

E. Costs

Dollars

Time

Purpose and Planning

Shepard identified the important issues concerning planning--i.e., its impact on data collection and reporting and the distinction between formative and summative evaluation. However, it is essential to go beyond that to carefully identify the key decisions for which evaluation data are needed.

Unless key decisions are identified, evaluation (especially formative) may be misdirected in one of two ways. First of all, it may be conducted strictly in a reactive or narrowly focused fashion. That is, whenever a particular issue becomes of great concern, e.g., when funds are about to be allocated by the legislature, a quick survey is taken of the opinions of, say, fourth grade teachers. It may be done by someone not intimately knowledgeable about the assessment and data are either superficial or inefficiently gathered. Whenever a particular concern or question arises one would do well to consider what other decisions are pending for which this data source could provide information and how this data gathering effort should be tied in with other such activities. At the other extreme is the case where someone attempts to gather detailed information on many aspects of the program just because the data are there. The Shepard checklist helps reduce the first type of error but unless key decisions (or some other criteria) are identified, it can easily promote the second type of error.

Thus, when planning an evaluation, some criteria are needed to limit the scope and direct the focus of the evaluation effort. I would contend that a useful criterion is a list of key decisions for which evaluation data are needed.

When developing the list of key decisions it is also very important to identify the people who will be making the decisions. Such people need to be involved in planning to ensure that they will find the data useful in decision making. It may be necessary on the part of the people doing the evaluation to argue for the importance of certain kinds of data to ensure that decisions are indeed based on an adequate range of information.

The activities which our team carried out could very well be considered as part of the planning stage for an evaluation. When we were approached, we were told that MEAP wanted formative evaluation but beyond that it was not specified what was needed. Thus MEAP used the approach of getting outsiders in to help plan any evaluation effort rather than plunge into either gathering information just because it was there or just zeroing in on a few issues which were a result of the "squeaky wheel" phenomenon.

Had the MEAP approached us with a list of decisions with which they or other groups (e.g., the legislature) were faced, I suspect the work of the panel would have been very different. Had such a list been presented, it could have been used as the basis for modifying the checklist and limiting the scope of the panel's work. I do not consider the lack of such a list being presented to us as necessarily a mistake on their part as long as they have the time and resources to take this more lengthy approach. The approach taken provides them with an external view of the key decisions which should guide future evaluation activities.

Staffing

Shepard emphasizes the wisdom of a team approach and the importance of determining the biases of team members. She makes some very good points in this regard. As part of our planning meeting in April, each panel member

completed a brief questionnaire (prepared by Shepard) to get at the biases of the team members. I will not discuss the results other than to say that no extreme biases existed according to this self-report measure. The results essentially confirmed the positions already assumed to be true by those who selected the panelists.

In regard to the use of such a measure to determine bias, I would suggest that it be used earlier in the selection process to identify a pool of potential team members.[8] In our case, the panelists were already selected and perhaps an additional person could have been added had some unexpected biases appeared, but otherwise little adjustment could be made.

I would like to add some comments about staffing to those made by Shepard. First, our experience suggests that it is very important to determine that the staff collectively has the ability to assess the quality of all aspects of assessment which are to be considered and that their qualifications are balanced with the purpose and magnitude of the evaluation effort. The team will be constituted differently both in size and expertise depending on whether they are doing the type of activity our panel did or if they are carrying out a more extensive data gathering and evaluation task in specified areas. Thus, a key issue in staffing becomes matching the staff configuration to the purpose of the evaluation activity and the resources to be devoted to the effort.

I would lean toward the approach of clearly defining the purpose of the evaluation effort, resources available, and biases of potential staff before any staffing decisions are made. (One of the problems was that the task was redefined after the team was assembled.)

---

[8]The identification of such a pool may be useful to other states as well.

Another important issue in regard to staffing is considering other approaches besides the site-visit-expert-team approach. The third party involvement is critical. However, the involvement of internal staff can be critical to ensuring eventual data utilization and it may also be more efficient and economical. If such an approach is used, however, care must be taken to define the rights and roles of the external parties to ensure objectivity in the evaluation. The hiring of data collectors who may not be highly trained also needs to be considered when staffing the evaluation.

Data Collection

Shepard has emphasized the need to consider alternative ways to gather data--document review, surveys, public hearings and interviews. Whole courses are devoted to such topics. The only comment I'd like to add as a result of our work with MEAP concerns the specificity of the questions asked and their relationship to the intended use for the data. For example, one survey we saw asked teachers how useful they found the results for classroom instructional planning.[9] The possible responses ranged from very useful to not at all useful. Another question went farther--"what impact has the student assessment results had on you?". Five ways data could have impact were given as options and teachers could add more categories. But the teachers were never asked how frequently they used the data or if they provided significant new insights beyond those provided by already existing information to help the teacher plan instruction. For example, the survey mentioned above showed that 56 percent of the fourth grade teachers who resporded found the results useful in confirming "previous

[9]Aquino, A. Teachers' Attitudes Regarding the Use of the 1974-75 Michigan Educational Assessment Program Test Results, July 1975.

tentative judgments about individual student skills which I had." But did teachers do that just for a couple students? Did they need such confirmation?

It is essential to carefully analyze the questions asked in any data collection effort in light of the uses to which the evaluation data are to be put. (What I'm suggesting is, in essence, the evaluation counterpart of objective-referenced testing.)

Data Sources

Shepard did not directly address the issue of data sources in her evaluation method. Although data sources are in many cases quite obviously defined by the topic of investigation, it may be helpful to treat it as a separate issue to see that the evaluators have indeed tapped all relevant sources on particular topics. For example, when considering goals there is a wide range of possible people who could serve as data sources for evaluating the adequacy of the goals. Who should be the sources--the legislature, teachers, state curriculum specialists, the superintendent of public instruction, the state board of education?

In the effects section of the checklist, Shepard lists numerous groups who may be affected and suggests these as sources of information. I agree that affected groups should be primary data sources when looking at effects, but when looking at other areas in the checklist other criteria may be needed for determining data sources. For example, (1) expertise in areas such as testing, sampling, budgeting and/or goal setting, or (2) a person's particular role or responsibility may determine whether or not to include someone as a data source.

Management

The management of an evaluation effort is often slighted in discussions of how to conduct an evaluation. Be it a short term effort such as the work

of our panel, or a full scale long term evaluation, a clear designation of responsibilities needs to be made. The approach used by the panel was that one person in MEAP was designated as the contact person with the panel. She met with the panel in April to come to agreement on their rights and responsibilities, which were then delineated in a memorandum of agreement. The MEAP contact person also sent the panel the documents which addressed their questions and made arrangements for the interviews requested by the panel. One person on the team was designated as team leader to coordinate team activities. Each team member was assigned specific questions to address and was given specific writing assignments for the final report.

The process worked quite well and responsibilities of each person were quite clearly delineated. Such clear delineation of responsibilities is essential, especially when the team is scattered across the country. When an even larger evaluation effort is being conducted, it is even more essential that careful consideration be given to the specific issues of management of the evaluation.

When making decisions about staffing, consideration needs to be given to management since it is necessary to have such a responsibility clearly defined. Also, factors such as the evaluation experience of each team member will affect the amount of time needed for management. For example, if some team members are not experienced evaluators, the person responsible for the team management may need to spend extra time clarifying tasks with those individuals.

Data Snythesis and Reporting

Based on our experience in Michigan, I would recommend that an evaluation method provide specific guidance to evaluators on data synthesis and reporting methods. Too often reports merely present a series of findings

in terms of significant differences or percentages of people holding different perspectives without tying them to any practical steps in terms of decision making. At times this may be appropriate but I fear that evaluators often stop short in terms of the interface with users of evaluation data. Evaluations, just as assessment programs, are made worthwhile when their results are used in decision making.

In terms of data synthesis, the approach used by our panel was to let each person use his or her own judgment in determining the areas of strengths and weaknesses regarding one section of the checklist. At the end of the interview days, the panel members each wrote down their major findings and we met together to review them to see if another panel member had contradictory information. (In most cases, we had had two panelists attend each interview, which helped provide a check on our findings.) Another check occurred when all panelists reviewed the entire report before its submission. Any point on which the team could not reach consensus was eliminated from the report or a minority view also was presented. Before the report was formally presented to the MDE Research, Evaluation and Assessment Services Director, a draft of the report was reviewed by an MEAP staff member for factual accuracy. The panel, of course, maintained the right to determine what was finally put in the report.

How evaluators report their findings involves more than just the format of a written document. In our case the reporting agreement we had with MEAP was that we would present a draft report to the MEAP Supervisor and Research, Evaluation and Assessment Services Director by June 15, an oral presentation on June 16 and the final report by August 1, 1977. The panel was solely in charge of developing and editing the final report. MEAP staff may write and disseminate any separate statement (such as an endorsement, a rebuttal, a commentary or a descriptive piece). Should the Director decide to disseminate

a document describing the report, it will be identified as being different from the panel's report. Only the panel members' final report as prepared by them will be distributed with their names on it. In addition, the release of information from the report prior to MDE dissemination of it must have prior approval of the Research, Evaluation and Assessment Services Director except aspects which deal with the evaluation of the Shepard Checklist. All of these precautions in reporting are important ones to have in the agreement between the evaluators and the assessment program to protect the rights of both parties.

Unfortunately, none of the agreements on reporting were designed specifically to promote data use in decision making. In our report we point out that MEAP needs to strengthen its communication efforts so that data are used in decision making. The same type of criticism could be levied against us as an evaluation team if we do not actively promote, through our report and in conversation with MEAP staff, the importance of being sure that the appropriate decision makers are presented with the information in the report. It may have been wise for the panel to have requested that provision be made in the initial agreement for a followup meeting with the key decision makers to whom our report was directed. Such a meeting could help to ensure that the data have indeed reached the appropriate people and are being used to reach decisions on next steps in terms of carrying out the recommendations or presenting a basis for another course of action. Another possibility would have been for the panel to request an agreement from the program director that the staff would respond verbally or in writing concerning how they are going to deal with the recommendations.

Some may say that the evaluator has no right to make such a request of the "employer." But, if an evaluator is doing the job for more than the monetary payment and wants to be sure that the time spent was not wasted,

such agreements may help to increase the use of evaluation results. It may also be useful to the client in getting segments of an organization together and communicating about common concerns. We all know how easily various groups within an organization can become separated and need an outside impetus to re-establish communication channels.

SUMMARY

In this presentation I have elaborated on the Shepard method of evaluating an assessment program. The discussion addressed the Shepard checklist of topics to focus on when evaluating an assessment program as well as the following topics:

- purpose and planning
- staffing the evaluation
- data collection
- data sources
- management
- data synthesis and reporting

Generally, the points made are equally applicable to evaluations other than those related to assessment programs.

APPENDIX A

A REVIEW OF THE MICHIGAN EDUCATIONAL ASSESSMENT PROGRAM (MEAP)

Accountability

The Michigan accountability model is actually a planning model for rational decision making. Each step of the model follows logically from the preceding step, and each generates the activities for the next step. The six steps of the model are:

1) The Identification of Common Goals
2) The Development of Performance Objectives
*3) Assessment of Needs
4) Analysis of Delivery Systems
5) Evaluation
6) Recommendations for Improvement

The assessment program is a part of the six-step accountability model developed by the Michigan Department of Education. Though the assessment program began prior to the completion of the first two steps of the model, it is now an integral part of the accountability process.

The accountability concept is based on the assumption that informed people will make better decisions. Accountability requires that all participants in the educational process become informed about how the system is working and use that information to improve the system.

Educational accountability requires the gathering and reporting of information on what educators are trying to accomplish, how well they are doing, what methods are being used with what success and at what costs, and what changes are recommended. Accountability can be used by local educators at the building and district level or at the state level to evaluate the overall status and progress of the state's educational efforts. Only by periodically gathering systematic information can areas of weakness be improved and areas of strength maintained.

History

The Michigan Educational Assessment Program (MEAP) since January, 1970, has provided information which contributes to the assessment of educational needs. The Program was initiated by the State Board of Education, supported by the Governor, and funded by the Legislature initially through enactment of Act 307 of the Public Acts of 1969, and subsequently under Act 38 of the Public Acts of 1970.

Norm-Referenced Tests

The first four years of the assessment program (1969-1973) made use of standardized, norm-referenced testing instruments to assess student performance in vocabulary, reading comprehension, English usage, and arithmetic. In addition, attitudinal and socioeconomic data were gathered in certain years.

---

*Prepared by Martha S. Caswell of the Michigan Department of Education, March 1977.

However, it was decided that the information provided by the norm-referenced tests did not adequately serve the purpose of MEAP to provide information to the State Board, the Executive Office, the Legislature, local educators, teachers, students, and parents on the status and progress of Michigan basic skills education. Therefore, the norm-referenced tests were replaced by objective-referenced tests in the 1973-74 school year.

Objective-Referenced Tests

In accordance with Michigan's educational goals, a series of learning areas were selected for assessment at five grade levels (1, 4, 7, 10, and 12):

Preprimary (objectives for entry to Kindergarten and grade 1)

Art
Communication Skills
Foreign Languages
Health Education
Mathematics and Metric Measurement
Music
Physical Education
Science
Social Studies

Minimum objectives for end of grades 3, 6, and 9)

Life Role Competencies (Objectives for end of Grade 12)

Objectives to be Tested

The minimal skill performance objectives for these areas were developed under the direction of staff from the Instructional Specialist Program of the General Education Services, and with the help of committees of Michigan educators and citizens. The State Board received and approved each set of objectives (except for the Life Role Competencies which are still in a developmental stage). Each set will be reviewed and revised periodically. These performance objectives do not cover all skills in each subject area but are, instead, minimal skill objectives which most students are expected to attain. Local school districts should supplement these objectives with others in order to provide pupils with the breadth and depth of educational experiences needed to achieve their potential.

Grades 1, 4, 7, 10, and 12 were selected as the five points at which to assess student performance because they represent important transition points in each student's education: the beginning of primary education, completion of primary education, completion of middle school education, and the completion of high school.

The Grade 12 Life Role Competencies are in a developmental stage at this point in time. The idea of using them as the basis for some kind of 12th grade exit test has been discussed, but as of now, no plan has been developed for assessing these objectives. How the objectives will be used depends in part on the characteristics of the objectives themselves.

Test Development

Test items have been written by Michigan educators in school districts which had contracts with the MDE or in item-writing conferences in Michigan. Try-outs and critiquing have been done under contracts with test development agencies.

The test development process uses a systems approach in which items are written, reviewed, revised, tried out, reviewed, revised, tried out, etc., until a group of appropriate items is available.*

1976-77 Program

1976-77 marked the eighth year of the Assessment Program and the fourth year of objective-referenced testing. Students were tested on an every-pupil basis from September 13-October 8 in reading and mathematics at Grades 4 and 7.

Grades 4 and 7

Only a selected number of the minimal performance objectives in reading and mathematics are tested in each grade. At Grade 4, 19 reading objectives and 30 mathematics objectives are tested. At Grade 7, 20 reading and 40 mathematics objectives are tested.

Each objective is tested by a set of five items. A student correctly answering four or five items for an objective is judged to have attained that objective. Grade 4 assessment tests contain 245 items (95 reading items and 150 mathematics items). Grade 7 assessment tests contain 300 items (100 reading items and 200 mathematics items).

The tests are untimed. This allows students to work at their own pace without penalty. The Assessment Program is interested in determining whether or not students have attained selected basic skills, not in whether or not students can demonstrate these skills within a timed period. However, some students do not finish the tests because the time necessary to do so would result in a high degree of frustration and anxiety on the part of the student. Teachers are encouraged to use their own discretion when testing such students.

Experimental Tests

In addition to the "core" tests administered to fourth and seventh graders there are also experimental tests given on a sampling basis. These experimental tests are used in the process of test item try-outs and validation. Experimental tests are given in mathematics and reading and in other selected subject areas.

---

*For more information on the development of the assessment program, see The Long Range Plan of the Michigan Educational Assessment Program, Michigan Department of Education, Lansing, Michigan 48909.

The other selected subject areas tested in 1976-77 were music and metric measurement. In addition to providing statistical data on the behavior of test items, these tests also give an indication of how students are doing on a statewide basis in the subject areas selected. The two areas tested in previous years are health and science.

Grade 10

1976-77 was the second year that testing has been done on the tenth grade level on a voluntary basis. Some 170 schools volunteered to participate in the program this year. Selected objectives were tested on an every-pupil basis in the volunteer schools in the subject areas of reading and mathematics. There were four different test forms, two for reading and two for mathematics. Each objective was tested by a set of eight items, five items for reporting and three for item validation. In general, the reaction of participating schools to the tenth grade program has been positive.

Kindergarten

The Kindergarten testing program is also conducted on a volunteer basis. The Kindergarten program has been offered for 2 years now and contains several feature that are different from the testing program at Grades 1, 4, 7 and 10. Testing children in this age group requires the use of testing methods that differ from those used with older children (i.e., paper-and-pencil items).

The Kindergarten tests are based on the Preprimary Performance Objectives. Teachers are provided with the objectives, suggestions for ways of testing the objectives, and forms for recording student performance. Testing is done throughout the school year. Teachers may test any or all of the objectives and can use either the tests MEAP provides, their own tests, or a combination of the two. The Kindergarten assessment tests are provided as a service for the teachers and they are encouraged to use them in the way that will be most useful for them. Meetings are held periodically with Kindergarten teachers throughout the state to give teachers the opportunity to provide MEAP with feedback and suggestions for test improvement.

Grade 1

The first grade assessment program was introduced in the 1974-75 school year. Even though it is now three years old, it is still considered a pilot program. Again, because of the problems involved with administering a paper-and-pencil test to this age group, tests are administered on an individual basis or in a small group situation. This involves a good deal of teacher time so the Preprimary Objectives selected to be tested are few in number and the number of items per objective is also limited.

The tests are given on a sampling basis so that not all schools or first graders in Michigan are involved. The test results for the statewide sample yield reliable data on the objectives tested. This information is used to improve test items and administration procedures, for revising the objectives themselves, and for determining the extent and nature of educational needs as reflected by the results.

Test Reports

There are a number of test result reports for grades 4, 7, and 10, which are returned to teachers, principals, and district personnel. These report forms contain various kinds of information which are of interest to educators at different levels. The report forms provided are the Individual Student Report, the Classroom Listing Report, School Summary Report, District Summary Report, the Test Item Analysis, and the Proportions Report.*

A new feature of the Assessment Program this year was the Proportions Report provided to districts as part of the fourth and seventh grade test results. This Report indicates for a district and the schools within a district the percentage of students attaining each of four proportions of the mathematics or reading objectives.

Future Directions

Using the Proportions Report and three criteria, the state will identify schools with high, moderate, or low needs in reading and/or mathematics. Once high needs are identified, the state will work with a school which asks for help to ensure that students master the basic skills. This program is in the developmental stages at this point in time. The Department hopes to operationalize it in a few selected schools on a trial basis in June, 1977.

Pending action by the State Board of Education, tenth grade testing will be done on an every-pupil basis in reading and mathematics next year. The State Board should approve or disapprove this plan by April, 1977. If the Board approves it, the Assessment Program is ready to implement the program in the fall.

Next year Art and Social Studies tests will be administered on a sample basis. This will result in a total of six performance objective areas (other than reading and mathematics) which will have been tested on a statewide basis. After all seven subject areas (for which performance objectives have been developed) have been tested, the process will begin again.

The Assessment Program, because of recommendations from math specialists, will include several metric objectives as part of the grade 4 and 7 mathematics objectives tested next year. These metric objectives will be in addition to the regular math objectives tested each year. As the transition is made from the measurement system now used to the metric system, more metric objectives will be included on the assessment tests.

The Assessment Program, as part of the accountability model, will continue to assess the status and progress of Michigan basic skills education. The assessment data will be used to identify schools with low, moderate, and high needs in the areas of math and reading. The use of assessment data in this way will enable the state to work with schools to continue to improve the quality of education received by Michigan students.

---

*For those interested in a more detailed discussion of the report forms, see the Interpretive Manual, 1976-77, Michigan Department of Education, Lansing, Michigan 48909.

APPENDIX B

Specific Questions Derived from Shepard Checklist for Use with MEAP[1]

A. GOALS AND PURPOSES

Does sufficient evidence exist to judge the adequacy of . . .

1. the established goals?
2. consideration given to the following kinds of goals
   a. pupil diagnosis
   b. pupil certification
   c. program evaluation, improvement, selection
   d. resource allocation
   e. accountability
   f. research
   g. informing the public
3. the importance, uniqueness, feasibility and scope of the goals?
4. *the process used to establish goals*[2]
5. *the existing process to modify goals?*
6. *the clarity of understanding among affected groups concerning what the goals are?*

B. TECHNICAL ASPECTS

Does sufficient evidence exist to judge the adequacy of . . .

1. the basis for selecting norm-referenced or objective-referenced tests?
2. the consistency of goals to the choice of norm-referenced or objective-referenced tests?
3. the process used to select content areas for testing?
4. the process for development and review of content objectives?
5. the process used to gather input on objectives to be covered?
6. the process used to select content objectives to be tested?
7. the match of test items to content objectives?
8. how cultural bias is defined, identified, and/or eliminated?
9. the procedures used to field test items/tests?
10. *the procedures used to revise tests?*

---

[1]The questions are stated such that the adequacy of evidence on each topic could be assessed. Had the task been to actually judge the adequacy of each aspect of MEAP, the lead part of the questions could have been changed to read "Does the evidence indicate adequacy of . . ."

[2]Questions in italics correspond to italicized items on the checklist.

B. TECHNICAL ASPECTS (continued)

Does sufficient evidence exist to judge the adequacy of . . .

11. *the general process used for test development?*
12. the sampling procedure?
13. the test administration process and method of handling missing data?
14. *the conduct and use of item analysis?*
15. longitudinal analyses?
16. the usefulness of reporting variables?
17. reporting procedures (e.g., appropriateness of data analyses, use of different reports for different audiences, appropriateness of interpretation)?
18. *how credible audiences cond der the assessment data to be?*
19. *how well results are disseminated to intended audiences?*

C. MANAGEMENT

Does sufficient evidence exist to judge the adequacy of . . .

1. long and short term planning (e.g., use of PERT or other management schemes)?
2. timelines?
3. the planning process?
4. documentation of procedures?
5. ongoing formative evaluation?
6. attention to personnel issues, e.g.,
   a. composition of the total staff in areas of affirmative action, equal opportunity, technical skills?
   b. employee formal and informal evaluation?
   c. staff training and development?
   d. human relations?
7. procedures used to review support services (e.g., supplies, secretarial assistance)?
8. how subcontracting is done, e.g.,
   a. the process/basis used to decide to do subcontracting?
   b. the basis for selection of a subcontractor?
   c. the processes used to monitor subcontracting?
9. the internal communication process and decision-making procedures, e.g.,
   a. *how elements of the agency communicate?*
   b. how managerial and policy decisions are distinguished?

C. MANAGEMENT (continued)

Does sufficient evidence exist to judge the adequacy of . . .

10. procedures for redress of grievances by those affected by assessment?
11. the timelines of results reaching users?
12. *the division of responsibilities among staff (both within the assessment staff and between assessment staff and other groups working on tasks impinging on assessment)?*
13. *the allocation of time to various endeavors?*
14. *how assessment fits in with other state department of education activities?*
15. *use of advisory groups?*

D. INTENDED AND UNINTENDED EFFECTS

1. What additional documented information is needed from the affected groups (see Checklist) on effects (see Checklist)?
2. What evidence exists that MEAP has an adequate process for determining intended and unintended effects?
3. What evidence exists about the kind of information collected and how well MEAP uses it?
4. Does evidence exist on the following effects?[3]
   a. To what extent and specifically how do teachers, administrators, parents, legislators and others use the results?
   b. What effects does the test taking have on children?
   c. What uses are made of the test results which are not intended?
   d. How has the assessment affected the curriculum of those areas assessed at local and state levels?
   e. How has assessment affected the curriculum in areas not tested at local and state levels?
   f. How has assessment affected teaching at grade levels other than the ones at which testing occurs?
   g. What effect has assessment had on school and district testing programs?
   h. To what extent have stated goals been attained?
   i. What differences exist in terms of immediate and long term effects?
   j. How have people's opinions changed about the educational process as a result of receiving assessment data?

[3]The questions listed in this category are ones explicated after the planning meeting and appear especially important for Michigan to investigate.

D. INTENDED AND UNINTENDED EFFECTS (continued)

4. Does evidence exist on the following effects? (continued)

   k. What effect has assessment data on the existence of an assessment program had on legislative actions?

   l. What detrimental uses have been made of data?

   m. What detrimental uses do people fear will be made of data?

   n. How has assessment data been used to make resource allocations?

   o. Could the decisions made based on assessment data been made as well with other already existing information?

   p. Has public confidence in education increased as a result of having access to data?

   q. Has the public become more interested in education as a result of having access to data?

   r. What proposed changes in assessment procedures are most acceptable to various audiences? (Expected effects)

E. COSTS

Does sufficient evidence exist to judge the adequacy of . . .

1. the dollar costs for the total program, including state and district level costs?
2. the time costs to students, teachers and administration?
3. the utility of the cost information?